Sparse Canonical Correlation Analysis via Truncated ℓ_1-norm with Application to Brain Imaging Genetics

Lei Du*, Tuo Zhang*, Kefei Liu†, Xiaohui Yao†, Jingwen Yan†,
Shannon L. Risacher†, Lei Guo*, Andrew J. Saykin† and Li Shen†§ for the ADNI
*School of Automation
Northwestern Polytechnical University, Xi'an, China 710072
Email: dulei@nwpu.edu.cn
†Indiana University School of Medicine, Indianapolis, USA 46202
§Corresponding to: Email: shenli@iu.edu

***Abstract*—Discovering bi-multivariate associations between genetic markers and neuroimaging quantitative traits is a major task in brain imaging genetics. Sparse Canonical Correlation Analysis (SCCA) is a popular technique in this area for its powerful capability in identifying bi-multivariate relationships coupled with feature selection. The existing SCCA methods impose either the ℓ_1-norm or its variants. The ℓ_0-norm is more desirable, which however remains unexplored since the ℓ_0-norm minimization is NP-hard. In this paper, we impose the truncated ℓ_1-norm to improve the performance of the ℓ_1-norm based SCCA methods. Besides, we propose two efficient optimization algorithms and prove their convergence. The experimental results, compared with two benchmark methods, show that our method identifies better and meaningful canonical loading patterns in both simulated and real imaging genetic analyse.**

***Keywords*-Sparse Canonical Correlation Analysis, Truncated ℓ_1-norm, Brain Imaging Genetics**

I. Introduction

Brain imaging genetics has gained more and more attentions recently [1], [2]. A major task of imaging genetics is to identify bi-multivariate associations between single nucleotide polymorphisms (SNPs) and imaging quantitative traits (QTs). Sparse canonical correlation analysis (SCCA), which is powerful in bi-multivariate relationship discovery coupled with feature selection, has become a popular technique in imaging genetic studies [3], [4], [5], [6], [7].

Witten *et al.* [3] introduced the ℓ_1-norm (Lasso) to assure sparsity which only selects a small proportion of the features. Since then, many SCCA methods using the ℓ_1-norm or its variants are proposed [8]. There are two major concerns regarding them. First, the ℓ_0-norm, which only penalizes those nonzero features, is the most ideal constraint. But it is neither non-convex nor discontinuous [9]. Second, the ℓ_1-norm constraint is not a stable feature selector and thus could incur estimation bias [10].

To overcome the problem above, the truncated ℓ_1-norm penalty (TLP) [10], [11] is proposed. The TLP is defined as $J_\tau(|x|) = \min(\frac{|x|}{\tau}, 1)$ with τ being a positive tuning parameter. It approximates ℓ_0-norm and permits desirable sparsity. In addition, TLP can be equivalently transferred to a piecewise linear function, and thus is easy to handle.

In this paper, we propose the TLP based SCCA (TLP-SCCA) which embraces the TLP into the CCA model. The TLP-SCCA has the following advantages [10]. First, the TLP performs as a tradeoff between the ℓ_0 and ℓ_1 functions. This means that it not only has improved feature selection, but also can be solved effectively. Second, it is an adaptive shrinkage method if τ is tuned appropriately. We propose two effective optimization algorithms, both using the alternating direction method of multipliers (ADMM) technique [12], and they are guaranteed to converge. The experimental results, compared with two popular ℓ_1-norm based SCCA [3], [6], show that both TLP-SCCA exhibit cleaner canonical loading patterns than the ℓ_1-SCCA.

II. The Truncated ℓ_1-norm Penalty

In this paper, a boldface lowercase letter denotes a vector, and a boldface uppercase letter denotes a matrix. $\mathbf{X} \in R^{n\times p}$ denotes the SNP data, and $\mathbf{Y} \in R^{n\times q}$ is the QT data.

The truncated ℓ_1-norm is defined as follows [13]:

$$P^{TLP}(\mathbf{u}) = \sum_i J_\tau(|u_i|), \quad \text{where} \quad J_\tau(|u_i|) = \min(\frac{|u_i|}{\tau}, 1). \tag{1}$$

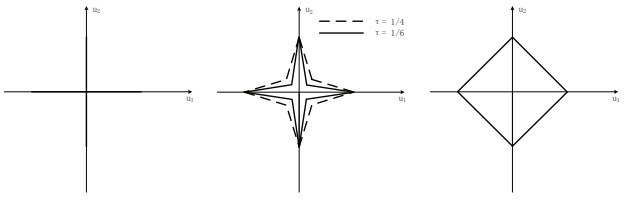

Figure 1. Visualization of the ℓ_0-norm ball (left), TLP ball with $\tau = \frac{1}{4}$ and $\tau = \frac{1}{6}$ (middle), and ℓ_1-norm ball (right).

The parameter τ is a threshold. Given an appropriate τ, TLP balances between the ℓ_0-norm and ℓ_1-norm according to the magnitude of the coefficients. Fig. 1 presents the norm ball of ℓ_0-norm, ℓ_1-norm, and TLP with different τ's. The

This is the author's manuscript of the article published in final edited form as:
Du, L., Zhang, T., Liu, K., Yao, X., Yan, J., Risacher, S. L., . . . Shen, L. (2016). Sparse Canonical Correlation Analysis via truncated ℓ1-norm with application to brain imaging genetics. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). http://dx.doi.org/10.1109/bibm.2016.7822605

TLP ball is more pointy than the ℓ_1 ball, and thus it is more desirable in sparse learning.

Since TLP is non-convex, we employ the Difference of Convex functions (DC) programming technique [14], [11]. Let $f_1(u_i) = |u_i|$ and $f_2(u_i) = \max(|u_i| - \tau, 0)$, we have:

$$J_\tau(|u_i|) = \frac{1}{\tau}(f_1(u_i) - f_2(u_i)) \tag{2}$$

where both $f_1(u_i)$ and $f_2(u_i)$ are convex. The affine minorization of $f_2(\cdot)$ can be denoted as:

$$f_2(u) = f_2(u_i) + < u - u_i, d_i >, \tag{3}$$

where $d_i = sign(u_i)\mathbb{I}(|u_i| > \tau)$ is the sub-gradient of f_2 at u_i with the $\mathbb{I}(\cdot)$ being an indicator function. Then the minimization of P^{TLP} can be transferred to its subproblem,

$$\arg\min P^{TLP} = \arg\min \frac{1}{\tau}(f_1(\mathbf{u}) - \mathbf{D}\cdot\mathbf{u}), \tag{4}$$

where $\mathbf{D}$ is a diagonal matrix with the i-th element as d_i.

III. The Novel SCCA via Truncated ℓ_1-norm

A. The TLP-SCCA Model

The TLP-SCCA can be formally defined as follows:

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \tag{5}$$

$$s.t.\ \ ||\mathbf{X}\mathbf{u}||^2 \le 1, ||\mathbf{Y}\mathbf{v}||^2 \le 1, P_1^{TLP} \le c_1, P_2^{TLP} \le c_2,$$

where $P_1^{TLP} = \sum_i J_\tau(|u_i|)$ and $P_2^{TLP} = \sum_j J_\tau(|v_j|)$ are the TLP penalties. The merits are three-fold. Firstly, it has lower estimation bias than a traditional ℓ_1-SCCA since it approximates the optimal ℓ_0-SCCA if given an appropriate τ. Secondly, it has better sparsity-inducing ability than the ℓ_1-SCCA. Thirdly, it is easier to solve than the ℓ_0-SCCA.

B. The Optimization Algorithm

The TLP-SCCA is biconvex in $\mathbf{u}$ and $\mathbf{v}$, and we solve it based on the Alternate Convex Search (ACS) strategy [15]. We show how to employ the ADMM algorithm to updating $\mathbf{u}$ with $\mathbf{v}$ fixed. [1] The ADMM combines the advantages of dual decomposition and augmented Lagrangian method [12], which divides a large global problem into a series of small local subproblems [11]. Introducing a slack variable $\mathbf{z}$, TLP-SCCA with respect to $\mathbf{u}$ conveys to:

$$\arg\min_{\mathbf{u}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\beta}{2}||\mathbf{X}\mathbf{u}||^2 + \frac{\lambda}{\tau}||\mathbf{z}||_1 - \frac{\lambda}{\tau}\mathbf{D}\cdot\mathbf{u} \tag{6}$$

$$s.t.\ \ \mathbf{u} - \mathbf{z} = \mathbf{0}$$

where β and λ are tuning parameters, τ is the parameter of the TLP, and $\mathbf{D}$ is a diagonal matrix with the i-th element as the sub-gradient $sign(u_i)\mathbb{I}(|u_i| > \tau_1)$ (Recall Section II).

[1] For ease of description, we only show how to solve $\mathbf{u}$ with $\mathbf{v}$ fixed as a constant. The $\mathbf{v}$ can be solved via swapping $\mathbf{u}$ and $\mathbf{v}$.

Algorithm 1 The Truncated ℓ_1-norm Penalty based SCCA (TLP-SCCA)

Require:
$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^T$, $\Gamma \succeq 0$, $\mathbf{z} = \mathbf{0}$, $\mu = \mathbf{0}$
Ensure:
Canonical loadings $\mathbf{u}$ and $\mathbf{v}$.
1: Initialize $\mathbf{u} \in R^{p\times 1}$, $\mathbf{v} \in R^{q\times 1}$;
2: Compute Cholesky factorization $\mathbf{F_u}$ and $\mathbf{F_v}$;
3: **while** not converged **do**
4: Solve $\mathbf{u}$ according to Eq. (10);
5: Solve $\mathbf{z}_1$ according to Eq. (13);
6: Solve μ_1 according to Eq. (14);
7: Solve $\mathbf{v}$ in accordance with Step 4 $\sim$ 6 by swapping $\mathbf{u}$ and $\mathbf{v}$;
8: Scale $\mathbf{u}$ so that $||\mathbf{X}\mathbf{u}||^2 = 1$, and $\mathbf{v}$ so that $||\mathbf{Y}\mathbf{v}||^2 = 1$.
9: **end while**

Then we construct the augmented Lagrangian function:

$$\begin{aligned}\mathcal{L}_{\mathbf{v},\rho}(\mathbf{u},\mathbf{z},\mu) = &-\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\beta}{2}||\mathbf{X}\mathbf{u}||^2 + \frac{\lambda}{\tau}||\mathbf{z}||_1 \\ &-\frac{\lambda}{\tau}\mathbf{D}\mathbf{u} + \mu^{\mathbf{T}}(\mathbf{u}-\mathbf{z}) + \frac{\rho}{2}||\mathbf{u}-\mathbf{z}||^2\end{aligned} \tag{7}$$

where μ is the augmented Lagrangian multipliers, and $\rho \ge 0$ is the dual update step length. According to ADMM, this problem can be solved by sequentially updating $\mathbf{u}$, $\mathbf{z}$ and μ.

1) $\mathbf{u}$-update: In each iteration, $\mathbf{u}$ can be obtained from:

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\beta}{2}||\mathbf{X}\mathbf{u}||^2 - \frac{\lambda}{\tau}\mathbf{D}\mathbf{u} + \mu^{\mathbf{T}}(\mathbf{u}-\mathbf{z}) + \frac{\rho}{2}||\mathbf{u}-\mathbf{z}||^2 \tag{8}$$

where $\mathbf{z}$, μ and $\mathbf{v}$ are all fixed as constant.

Then we can solve $\mathbf{u}$ by $\mathbf{u} = \mathbf{F}^{-1}\mathbf{b}$, where

$$\mathbf{F} = \beta\mathbf{X}^T\mathbf{X} + \rho\mathbf{I}, \ \ \mathbf{b} = \mathbf{X}^T\mathbf{Y}\mathbf{v} - \mu + \rho\mathbf{z} + \frac{\lambda}{\tau_1}diag(\mathbf{D}). \tag{9}$$

Obviously, $\mathbf{F}$ remains unchanged during the iteration, and thus we compute it at the beginning of the algorithm. The Cholesky factorization of the positive definite $\mathbf{F}$ is $\mathbf{F} = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R}$ is an upper triangular matrix. Then $\mathbf{u}$ can be obtained by solving two linear system equations:

$$\hat{\mathbf{u}} = (\mathbf{R}^T)^{-1}\mathbf{b}, \ \ \tilde{\mathbf{u}} = \mathbf{R}^{-1}\hat{\mathbf{u}}. \tag{10}$$

2) $\mathbf{z}$-update: Likewise, $\mathbf{z}$ can be obtained from:

$$\tilde{\mathbf{z}} = \arg\min_{\mathbf{z}} \frac{\lambda}{\tau}||\mathbf{z}||_1 + \mu(\tilde{\mathbf{u}}-\mathbf{z}) + \frac{\rho}{2}||\tilde{\mathbf{u}}-\mathbf{z}||^2, \tag{11}$$

We equivalently reformulate the equation as follows.

$$\tilde{\mathbf{z}} = \arg\min_{\mathbf{z}} \frac{\rho}{2}||\tilde{\mathbf{u}} + \frac{\mu}{\rho} - \mathbf{z}||^2 + \frac{\lambda}{\tau}||\mathbf{z}||_1 \tag{12}$$

Then the soft-thresholding method can be used:

$$\tilde{\mathbf{z}} = S_{\lambda/(\rho\tau)}(\tilde{\mathbf{u}} + \frac{\mu}{\rho}), \ \ \text{where} \ \ S_\lambda(a) = sign(a)\max(|a| - \lambda, 0) \tag{13}$$

3) μ-update: The dual update of μ is shown as follows.

$$\tilde{\mu} = \mu + \rho(\tilde{\mathbf{u}} - \tilde{\mathbf{z}}). \tag{14}$$

According to the ACS strategy, $\mathbf{v}$ can be solved similarly by fixing $\mathbf{u}$. The key steps are exhibited in Algorithm 1.

Algorithm 2 The TLP-SCCA with Restart (rTLP-SCCA)

Require:
$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^T$, $\Gamma \succeq 0$, $\mathbf{z} = \mathbf{0}$, $\mu = \mathbf{0}$, $\eta \in (0, 1)$
Ensure:
Canonical loadings $\mathbf{u}$ and $\mathbf{v}$.

1: **while** not converged **do**
2: Compute $\mathbf{u}$, $\mathbf{z}_1$ and μ_1 using Algorithm 1;
3: **if** $r_c^k < \eta r_c^{k-1}$ **then**
4: $\alpha^{k+1} = \frac{1+\sqrt{1+4\alpha^{k\,2}}}{2}$, $\mathbf{z}_1^{k+1} = \mathbf{z}_1^k + \frac{\alpha^k - 1}{\alpha^{k+1}}(\mathbf{z}_1^k - \mathbf{z}_1^{k-1})$;
5: $\mu_1^{k+1} = \mu_1^k + \frac{\alpha^k - 1}{\alpha^{k+1}}(\mu_1^k - \mu_1^{k-1})$;
6: **else**
7: $\alpha^{k+1} = 1$, $\mathbf{z}_1^{k+1} = \mathbf{z}_1^{k-1}$, $\mu_1^{k+1} = \mu_1^{k-1}$, $r_c^k = \frac{r_c^{k-1}}{\eta}$;
8: **end if**
9: Solve $\mathbf{v}$ in accordance with Step 4 $\sim$ 8 by swapping $\mathbf{u}$ and $\mathbf{v}$;
10: Scale $\mathbf{u}$ so that $||\mathbf{Xu}||^2 = 1$, and $\mathbf{v}$ so that $||\mathbf{Yv}||^2 = 1$.
11: **end while**

C. The Optimization Algorithm with Restart

We now propose an optimization algorithm with more stability. The algorithm uses a restart rule depending on the summation of the running primal and dual error [16].

We define the primal residual at the $(k+1)^{th}$ iteration $r_p^{k+1} = \mu^{k+1} - \mu^k = \rho(\mathbf{u}^{k+1} - \mathbf{z}^{k+1})$, and the dual residual as $r_d^{k+1} = \mathbf{z}^{k+1} - \mathbf{z}^k$. Then the combined residual is

$$r_c^{k+1} = \frac{1}{\rho}||\mu^{k+1} - \mu^k||^2 + \rho||\mathbf{z}^{k+1} - \mathbf{z}^k||^2. \tag{15}$$

The key steps of the rTLP-SCCA (TLP-SCCA with restart) are shown in Algorithm 2. If the residual r_c decreases by a factor of at least η (Step 3), we use the Nesterov's accelerated gradient descent method to accelerate [12] (Steps 4-5). Otherwise, we restart the iteration (Step 7).

D. Computational Analysis and Convergence

In Algorithm 1, the initialization step is simple. Step 2 of Cholesky factorization only needs to be done once. According to Eq. (10), Steps 4 involves solving two linear systems. Thus its time complexity is $O(pq)$. Step 5 is the soft-thresholding operator whose time consumption is $O(p)$. Step 6 is quite simple and its time complexity is $O(p)$. Thus the time complexity of Steps 4-6 is $O(pq)$. Similarly, the time complexity of solving $\mathbf{v}$ is $O(pq)$. In summary, the complexity of each iteration is $O(pq)$.

The difference between TLP-SCCA and rTLP-SCCA is the restart steps which are easy to calculate. Thus the time complexity of each iteration of rTLP-SCCA is also $O(pq)$.

The TLP-SCCA problem is biconvex in terms of $\mathbf{u}$ and $\mathbf{v}$. According to Theorem 4.5 in [15], the TLP-SCCA converges monotonically depending on two conditions: (1) The set $\mathbf{B} = (\mathbf{u}, \mathbf{v}) \in R^p \times R^q$ is bounded. (2) Both $\mathbf{u}$-update and $\mathbf{v}$-update are solvable.

We note that $\mathbf{u}$ is bounded by $||\mathbf{Xu}||^2 = 1$, and $\mathbf{v}$ is bounded by $||\mathbf{Yv}||^2 = 1$. Therefore, $\mathbf{B}$ is bounded because that $\mathbf{u}$ and $\mathbf{v}$ are bounded. Obviously, the $\mathbf{u}$-update and $\mathbf{v}$-update are solvable guaranteed by ADMM method. Therefore, the TLP-SCCA converges during iteration. [15].

IV. Experimental Study

A. Benchmark Methods

In this paper, we mainly focus on the improved performance of SCCA using the TLP regularization. Thus we compare our method with the ℓ_1-SCCA, other than those structured SCCA methods [5], [17], [7], [18], [19]. We use two different implementations of ℓ_1-SCCA as benchmark. The first one is named as L1-SCCA which is contained in the Penalized Multivariate Analysis software toolkit [3], and the second one is named as L1-S2CCA by considering each group consists of only one covariate [6].

B. Parameter Tuning

Take the $\mathbf{u}$-update as an example, we have four parameters λ, β, τ, and ρ, and two unknown variables $\mathbf{z}$, μ to be decided in advance. $\mathbf{z}$ is a slack variable which is split from $\mathbf{u}$. We simply set $\mathbf{z} = \mathbf{0}$. μ is the dual variable and is set to $\mathbf{0}$ too. For the remaining parameters, we further reduce the computational burden based on some heuristic tricks. First, we set the dual update step size ρ to a conservative value to permit a good accuracy. Second, τ is a key parameter in TLP-SCCA, thus it is reasonable to use an $\mathbf{u}$-related τ. We set $\tau = E(|\mathbf{u}|) + T * \sigma(|\mathbf{u}|)$ and update it during the iteration, where $E(|\mathbf{u}|)$ and $\sigma(|\mathbf{u}|)$ are the *mean* and *deviation* of $|\mathbf{u}|$. T is a constant which controls τ, and further controls how many u_i's are penalized by ℓ_0-norm. According to normal distribution, $T = 1.96$ indicating that 95% of $|\mathbf{u}|$ are smaller than τ. This will yield a reasonable sparse result. At last, we only have two parameters, i.e. λ_1 and β_1, left. Likewise, this is the same situation for $\mathbf{v}$-update. Thus we have four parameters to be tuned and utilize the **nested** 3-fold cross-validation to tune them. In this study, we tune the parameters from the range of 10^θ with $\theta \in [-3, -2, -1, 0, 1, 2, 3]$. All methods utilize the same partition and run on the same platform to ensure a fair comparison. The stopping criteria is $|obj(\mathbf{u})^k - obj(\mathbf{u})_{k+1}|/obj(\mathbf{u})_k \leq \epsilon$, where $\epsilon = 10^{-5}$ is a desirable error bound in this paper.

C. Results on Synthetic Data

Two synthetic datasets are used in the experiment. Both datasets are generated by the following steps. 1) We first set up two vectors $\mathbf{u}$ with length p and $\mathbf{v}$ with length q separately. We then let most of elements of both $\mathbf{u}$ and $\mathbf{v}$ be zero. 2) We introduce a latent variable $\mathbf{z}$ with $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_{n\times n})$. 3) We generate $\mathbf{X}$ with each element $\mathbf{x}_i \sim N(z_i\mathbf{u}, \sigma)$, where σ stands for the standard deviation of noise. We also create $\mathbf{Y}$ with $\mathbf{y}_i \sim N(z_i\mathbf{v}, \sigma)$.

Fig. 2 shows the estimated canonical loadings. Both TLP-SCCA and rTLP-SCCA obtain cleaner canonical loading patterns than the competing methods. Our methods have smaller estimation bias regarding the canonical loadings. Table I shows results of the estimated correlation coefficients. We observe that three methods, i.e. the L1-SCCA, TLP-SCCA and rTLP-SCCA, perform similarly. The results

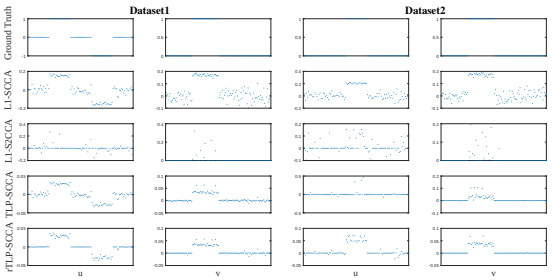

Figure 2. Canonical loadings estimated on synthetic datasets. The first two columns are for Dataset 1, and the second two columns are for Dataset2. The first row is the ground truth, and each remaining one corresponds to a method: (1) Ground Truth. (2) L1-SCCA. (3) L1-S2CCA. (4) TLP-SCCA. (5) rTLP-SCCA.

on synthetic datasets reveal that although TLP-SCCA and rTLP-SCCA do not hold the highest correlation coefficients, they obtain the most unbiased canonical loadings.

Table I
RESULTS ON SYNTHETIC DATA: THE ESTIMATED CORRELATION COEFFICIENTS OF EACH FOLD AND THEIR MEAN ARE SHOWN.

	Dataset 1	MEAN	Dataset 2	MEAN
Training Results				
L1-SCCA	0.99 0.99 0.98	0.99	0.98 0.99 0.98	0.98
L1-S2CCA	1.00 1.00 1.00	1.00	1.00 1.00 1.00	1.00
TLP-SCCA	0.99 0.99 0.99	0.99	0.98 0.97 0.98	0.98
rTLP-SCCA	0.99 0.99 0.99	0.99	0.98 0.98 0.98	0.98
Testing Results				
L1-SCCA	0.97 0.99 0.99	0.98	0.98 0.98 0.98	0.98
L1-S2CCA	0.74 0.85 0.89	0.83	0.92 0.82 0.91	0.89
TLP-SCCA	0.96 0.99 0.98	0.98	0.91 0.84 0.91	0.89
rTLP-SCCA	0.96 0.98 0.99	0.98	0.98 0.97 0.98	0.98

D. Results on Real Brain Imaging Genetics Data

We obtained the real imaging genetics data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of conducting ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to help detect the mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Please see www.adni-info.org for up-to-date information.

The genotyping and baseline amyloid imaging data of 283 non-Hispanic Caucasian participants were downloaded from the ADNI website. The imaging data were preprocessed in accordance with [7], and then the effects of the baseline age, gender, education and handedness were pre-adjusted by regression. After that we extracted 191 ROI level mean amyloid measurements. The genotyping data contain 58 SNP markers from the AD-related genes. Our goal is to verify if TLP-SCCA can find out the meaningful associations between the amyloid data and the SNP data.

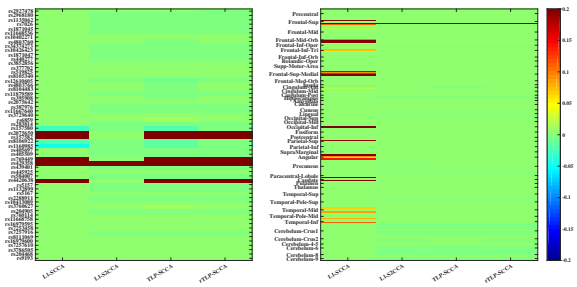

Figure 3. Heat map of estimated canonical loadings of different SCCA methods. The left panel shows the estimated weights of $\mathbf{u}$, and the right one shows the weights of $\mathbf{v}$.

Table II
RESULTS ON REAL IMAGING GENETICS DATA: THE ESTIMATED CORRELATION COEFFICIENTS OF EACH INDIVIDUAL FOLD AND THEIR MEAN ARE SHOWN.

Methods	Training Results	MEAN	Testing Results	MEAN
L1-SCCA	0.46 0.49 0.44	0.46	0.45 0.39 0.49	0.44
L1-S2CCA	0.47 0.51 0.46	0.48	0.48 0.43 0.51	0.47
TLP-SCCA	0.45 0.47 0.44	0.45	0.44 0.38 0.50	0.44
rTLP-SCCA	0.45 0.47 0.43	0.45	0.43 0.38 0.49	0.43

We first present the heat map regarding the estimated canonical loadings in Fig. 3. On the genotyping data (left penal), we clearly see that the L1-SCCA, TLP-SCCA and rTLP-SCCA find out 5 out of 58 SNPs. The five SNPs are rs429358 and rs769449 from *APOE*, rs157582 and rs2075650 from *TOMM40*, rs4420638 from *APOC1*. Note that all the five SNPs have been demonstrated to be highly correlated with AD. The L1-S2CCA only discovers the SNP rs429358, which means it loses the useful information carried by other four SNPs. On the imaging data (right panel), both TLP-SCCA and rTLP-SCCA identify only one strong signal, showing very clear and clean canonical loading pattern. The identified frontal region in the brain is known to be related with AD [20]. The results indicate that the TLP-SCCA can identify meaningful associations between genetic data and imaging data.

We also show the estimated correlation coefficients in Table II. There is no significant difference among various methods. This phenomenon reveals that generating a clean canonical pattern does not always correspond to a higher correlation coefficient [10]. Nevertheless, the TLP-SCCA and rTLP-SCCA have the potential to balance between the overpenalization and the overfitting, indicating their promising power in imaging genetic studies.

V. CONCLUSION

In this paper, we proposed the TLP-SCCA coupled with two effective optimization algorithms, i.e. TLP-SCCA and rTLP-SCCA. They identified cleaner canonical loading patterns which reduced the estimation bias of ℓ_1-SCCA. We analyzed their convergence and computation complexity.

We evaluated our methods and two ℓ_1-SCCA methods on both synthetic and real data. Both TLP-SCCA and rTLP-SCCA identified cleaner canonical loadings, indicating they successfully and accurately recognized the signals which were the closest to the ground truth. The results on the real imaging genetic data showed that TLP-SCCA and rTLP-SCCA yielded more meaningful canonical loading patterns than the benchmarks. Possible future directions include (1) investigating the performance trajectory under different τ's, and (2) using relevant structure information to guide the bi-multivariate association identification.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Funding: This research was supported by the National Natural Science Foundation of China (No. 61602384, No. 31500798 and No. 31671005), and Scientific Research Start-up Foundation (No. G2016KY0305) at Northwestern Polytechnical University. This work was also supported by NIH R01 EB022574, R01 LM011360, U01 AG024904 (http://adni.loni.usc.edu), RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG 042437, and R01 AG046171; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; NCAA 14132004; and CTSI SPARC Program.

REFERENCES

[1] S. Kim, S. Swaminathan, M. Inlow, S. L. Risacher, K. Nho *et al.*, "Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel," *PLoS One*, vol. 8, no. 7, p. e70269, 2013.

[2] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho *et al.*, "Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 792–814, 2015.

[3] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–34, 2009.

[4] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–34, 2009.

[5] X. Chen and H. Liu, "An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping," *Statistics in Biosciences*, vol. 4, no. 1, pp. 3–26, 2012.

[6] L. Du, J. Yan, S. Kim, S. L. Risacher, H. Huang, and et al., "A novel structure-aware sparse learning algorithm for brain imaging genetics," in *MICCAI*, 2014, pp. 329–336.

[7] L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher *et al.*, "Structured sparse canonical correlation analysis for brain imaging genetics: An improved graphnet method," *Bioinformatics*, vol. 32, no. 10, pp. 1544–1551, 2016.

[8] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer *et al.*, "Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers," *Brain imaging and behavior*, vol. 8, no. 2, pp. 183–207, 2014.

[9] G. Fung and O. Mangasarian, "Equivalence of minimal l 0-and l p-norm solutions of linear equalities, inequalities and linear programs for sufficiently small p," *Journal of optimization theory and applications*, vol. 151, no. 1, pp. 1–10, 2011.

[10] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 223–232, 2012.

[11] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, "Feature grouping and selection over an undirected graph," in *SIGKDD*. ACM, 2012, pp. 922–930.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] T. Zhang *et al.*, "Multi-stage convex relaxation for feature selection," *Bernoulli*, vol. 19, no. 5B, pp. 2277–2293, 2013.

[14] P. D. Tao *et al.*, "Duality in dc (difference of convex functions) optimization. subgradient methods," in *Trends in Mathematical Optimization*. Springer, 1988, pp. 277–293.

[15] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, no. 3, pp. 373–407, 2007.

[16] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast alternating direction optimization methods," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2014.

[17] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," in *AISTATS*, 2012.

[18] J. Chen, F. D. Bushman *et al.*, "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, vol. 14, no. 2, pp. 244–258, 2013.

[19] L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher *et al.*, "Structured sparse cca for brain imaging genetics via graph oscar," *BMC Systems Biology*, vol. 10, no. 3.

[20] D. T. Stuss, C. A. Gow, and C. R. Hetherington, ""no longer gage": frontal lobe dysfunction and emotional changes." *Journal of consulting and clinical psychology*, vol. 60, no. 3, p. 349, 1992.